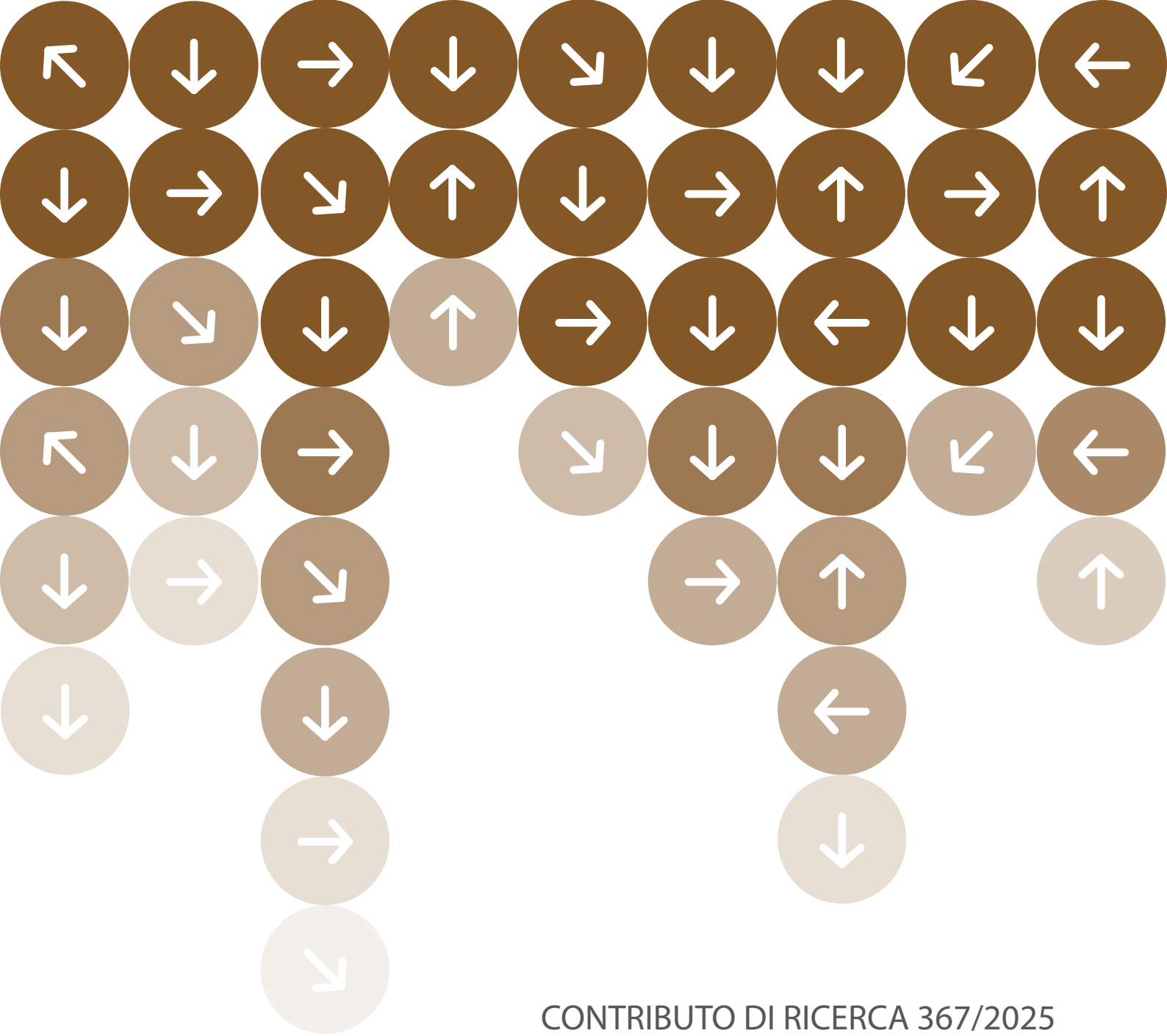


CONTRIBUTO DI RICERCA 367/2025

# DETERMINANTI DELLA PARTECIPAZIONE AL PSR 2014-2022 IN PIEMONTE: UN'ANALISI TRAMITE MACHINE LEARNING

Rapporto tematico realizzato nell'ambito dell'attività di valutazione del PSR 2014-2022 della Regione Piemonte

REGIONE PIEMONTE

PSR 2014-2022

IRES PIEMONTE

L'IRES PIEMONTE è un ente di ricerca della Regione Piemonte disciplinato dalla Legge Regionale 43/91 e s.m.i. Pubblica una relazione annuale sull'andamento socioeconomico e territoriale della regione ed effettua analisi, sia congiunturali che di scenario, dei principali fenomeni socioeconomici e territoriali del Piemonte.




Il documento in formato PDF è scaricabile dal sito www.ires.piemonte.it





via Nizza 18 – 10125 Torino – www.ires.piemonte.it

# DETERMINANTI DELLA PARTECIPAZIONE AL PSR 2014-2022 IN PIEMONTE: UN'ANALISI TRAMITE MACHINE LEARNING

RAPPORTO TEMATICO REALIZZATO NELL'AMBITO DELL'ATTIVITA' DI VALUTAZIONE DEL PSR 2014-2022 DELLA REGIONE PIEMONTE

## GLI AUTORI

Marco Adamo

# INDICE

# INTRODUZIONE

L'accesso ai finanziamenti pubblici rappresenta una leva cruciale per lo sviluppo e la competitività delle aziende agricole. Nell'ambito del Programma di Sviluppo Rurale (PSR) del Piemonte, tuttavia, la partecipazione delle imprese a tali misure non è sempre uniforme e risulta influenzata da una molteplicità di fattori economici, strutturali e geografici, sui quali è necessario indagare con l'obiettivo di fornire un supporto che aiuti i programmatori a identificare potenziali criticità nei meccanismi di delivery delle politiche.

La presente ricerca si propone di stimare i fattori principali che condizionano la partecipazione delle aziende agricole piemontesi ai bandi dell'Operazione 4.1.1 del PSR 2014-2022, impiegando modelli di Machine Learning (ML). In particolare, si utilizza l'algoritmo XGBoost per la modellazione predittiva, seguito da un'analisi interpretativa tramite SHAP (SHapley Additive exPlanations), al fine di identificare le variabili più influenti e comprenderne le interazioni.

I risultati di questa ricerca offrono spunti interessanti per la messa a punto delle politiche pubbliche. Infatti, comprendere i fattori che favoriscono o ostacolano la partecipazione ai bandi, permette di ottimizzare la progettazione delle future misure di finanziamento, rendendole più efficaci e accessibili. Grazie all'adozione di strumenti ML in questo contesto, è possibile superare i limiti dei tradizionali approcci statistici, fornendo analisi più robuste e dettagliate, in grado di supportare decisioni basate su evidenze empiriche. Le evidenze empiriche e metodologiche a supporto di queste asserzioni sono sviluppate nei capitoli successivi.

# MACHINE LEARNING E ANALISI DELLE POLITICHE PUBBLICHE

## L'IMPIEGO DI MACHINE LEARNING NELL'ANALISI DELLE POLITICHE PUBBLICHE

Negli ultimi anni, l'impiego del Machine Learning (ML) nell'analisi e valutazione delle politiche pubbliche ha guadagnato crescente attenzione accademica, grazie alla sua capacità di migliorare la previsione degli impatti delle politiche e la classificazione delle variabili rilevanti. Bell et al. (2022) hanno analizzato empiricamente il trade-off tra accuratezza e interpretabilità nei modelli ML applicati a contesti decisionali, dimostrando che algoritmi complessi come XGBoost[1], integrati con strumenti interpretativi come SHAP, possono raggiungere livelli di trasparenza comparabili a quelli di modelli lineari, superando i limiti dei metodi tradizionali. In ambito urbano, Zhou et al. (2024) hanno applicato XGBoost e SHAP per identificare i principali driver dell'innovazione nelle città cinesi, fornendo evidenze empiriche che supportano la pianificazione territoriale. Analogamente, Sha et al. (2024) hanno utilizzato XGBoost per classificare strumenti di politica sulla sicurezza alimentare in Cina, evidenziandone l'efficacia nell'individuare combinazioni di interventi ottimali.

Nel settore agricolo, il ML ha migliorato la comprensione delle dinamiche economiche e supportato la formulazione di politiche, superando le limitazioni dei modelli econometrici tradizionali. Ifft et al. (2018) hanno dimostrato che algoritmi come XGBoost offrono una maggiore capacità predittiva nella stima della domanda di credito agricolo, basandosi su dati di survey negli Stati Uniti, un approccio applicabile alla pianificazione dei finanziamenti agricoli. Shakoor et al. (2017) hanno utilizzato tecniche di ML supervisionato per prevedere la produzione agricola in Pakistan, mostrando come tali metodi possano informare le politiche di allocazione delle risorse. Gerber et al. (2024) hanno impiegato modelli come Random Forest per analizzare i divari di rendimento globale, identificando regioni a rischio di stagnazione produttiva e fornendo insight per politiche agricole sostenibili. Araújo et al. (2023) offrono una revisione sistematica dell'uso del ML in agricoltura, evidenziando il suo ruolo nella gestione dei raccolti e delle risorse, mentre Adnan et al. (2025) esplorano l'applicazione del ML alla gestione idrica, un aspetto cruciale per

[1] XGBoost (Extreme Gradient Boosting) è un algoritmo di machine learning basato su una tecnica di potenziamento (boosting) degli alberi decisionali. Si rimanda al paragrafo XXX dove viene spiegato in dettaglio.

l'ottimizzazione delle pratiche agricole. Queste evidenze metodologiche supportano l'applicazione del ML a contesti complessi come la partecipazione ai bandi del PSR Piemonte, offrendo una base solida per il presente studio.

# L'OPERAZIONE 4.1.1

L'Operazione 4.1.1 ha lo scopo di migliorare il rendimento globale delle aziende agricole sostenendo l'acquisizione, la costruzione, la ristrutturazione, l'ampliamento e la modernizzazione dei fabbricati e dei relativi impianti, nonché la dotazione di attrezzature e macchinari e l'impianto di coltivazioni legnose agrarie. Questo avviene tramite un contributo in conto capitale che cofinanzia gli investimenti.

Questa Operazione, in termini di risorse e partecipazione, è una delle azioni di policy più importanti nel caleidoscopio di misure, sotto-misure e operazioni che compongono il PSR del Piemonte. Infatti, secondo i dati presenti all'interno del Data Warehouse del PSR 2014 - 2022[2], le risorse stanziate complessivamente per questa Operazione ammontano a circa 165,5 milioni di euro: il 12% di tutte le risorse a disposizione del Piano Finanziario cofinanziato dal FEASR[3].

Nel corso della programmazione sono stati aperti 5 bandi (tab. 1) che hanno avuto un andamento variabile. Complessivamente si registrano 5.735 adesioni, indipendentemente dal buon fine della pratica. Questo numero equivale a circa il 14% di tutte le aziende piemontesi, il che è un buon indicatore del fatto che il settore agricolo ha un chiaro bisogno di fare investimenti in azienda.

**Tabella 1 - Domande presentate, ammesse e pagate sull'Operazione 4.1.1 per anno campagna**

| Anno campagna | 2015 | 2017 | 2019 | 2020 | 2021 | totale |
|---|---|---|---|---|---|---|
| **N. Beneficiari** | 2.040 | 931 | 1.152 | 595 | 1.017 | 5.735 |
| **Ammessi al finanziamento (operazioni)** | 838 | 372 | 248 | 414 | 399 | 2.271 |
| **Importo Ammesso al Finanziamento (euro)** | 107.150.475,19 | 64.861.638,21 | 27.077.339,13 | 22.297.493,04 | 70.599.563,52 | 291.986.509 |
| **Aiuto Ammesso al Finanziamento (euro)** | 44.221.352,48 | 26.922.553,61 | 11.583.728,88 | 8.918.997,22 | 29.297.582,80 | 120.944.215 |
| **Aiuto Pagato Totale (euro)** | 37.925.505,02 | 22.267.003,78 | 8.946.138,49 | 6.611.284,78 | 17.019.432,34 | 92.769.364 |

Fonte: Sistema Piemonte, Sviluppo Rurale – Data Warehouse

[2] http://www.sistemapiemonte.it/fedwanau/elenco.jsp
[3] La percentuale non tiene conto degli importi aggiuntivi derivanti da aiuti di Stato e EURI.

# METODOLOGIA

Il presente studio adotta un approccio quantitativo per valutare il contributo di variabili strutturali e geografiche alla probabilità di partecipazione delle aziende agricole piemontesi ai bandi dell'Operazione 4.1.1 del PSR 2014-2022.

Il dataset impiegato è costituito da tutte le aziende agricole attive presenti all'interno dell'Anagrafe Agricola Unica del Piemonte, per un totale di 40.904 osservazioni. Le variabili indipendenti selezionate sono l'età e il genere del titolare, la superficie agricola utilizzata (SAU), le unità di bestiame adulto (UBA)[4], il valore della Produzione Standard (PS)[5], l'Orientamento Tecnico Economico (OTE)[6] e la localizzazione geografica, secondo la classificazione territoriale del PSR. La variabile dipendente, ovvero il fatto che un'azienda abbia partecipato o meno a un bando, è dicotomica (variabile dummy) e assume valori "SÌ"/"NO". Questa variabile è stata creata utilizzando gli elenchi dei partecipanti ai bandi ed è stata unita al dataset delle variabili indipendenti tramite il Codice CUAA.

L'analisi è stata condotta utilizzando il software R Studio (versione 2024.12.1 build 563), utilizzando il pacchetto XGBoost (Chen & Guestrin, 2016) per la modellazione predittiva e treeshap (Lundberg & Lee, 2017) per l'interpretazione tramite SHAP values.

XGBoost (Extreme Gradient Boosting) costruisce una sequenza di alberi decisionali, dove ogni albero corregge gli errori del precedente tramite gradient boosting, minimizzando la differenza tra le previsioni e i valori reali. Il modello finale (F(x)) è espresso come:

$$F_M(X) = \sum_{m=1}^{M} f_m(x)$$

dove $f_m(x)$ è il contributo del singolo albero (m) ottimizzato mediante la funzione di perdita L(y, ŷ) (ad esempio, l'errore logaritmico per variabili binarie) con l'aggiornamento:

$$f_m(x) = argmin \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + f(x_i))$$

[4] L'UBA è l'unità di misura della consistenza di un allevamento che si ottiene applicando al numero dei capi presenti in azienda degli appositi coefficienti legati all'età ed alla specie degli animali. In tal modo si possono confrontare le dimensioni di allevamenti costituiti da specie diverse.

[5] La produzione standard (PS) calcolata in Euro è utilizzata per rappresentare il valore medio ponderato della produzione lorda totale di un'attività produttiva. Essa comprende sia il prodotto principale che eventuali sottoprodotti, esprimendo, quindi, il valore monetario della produzione agricola lorda "franco azienda".

[6] L'OTE è una variabile categorica che indica l'attività produttiva prevalente all'interno di una azienda. La determinazione dell'OTE avviene rapportando la PS risultante da ciascuna coltura o allevamento alla PS totale.

Per mitigare l'overfitting[7], cioè l'eccessivo adattamento del modello ai dati di addestramento, e migliorare la capacità predittiva, sono stati calcolati automaticamente i migliori parametri utili a regolare il modello, grazie all'impiego del pacchetto caret (Khun,2008).
L'addestramento del modello si è fermato alla 84ª iterazione, delle 200 impostate, grazie all'early stopping, un meccanismo che interrompe l'addestramento quando la performance sul set di test smette di migliorare, riducendo ulteriormente il rischio di overfitting.
Un'ulteriore ottimizzazione ha affrontato lo squilibrio dei dati: con le aziende non partecipanti circa 10 volte più rappresentate di quelle partecipanti, il parametro "scale_pos_weight", impostato a 9,92 (rapporto tra gruppi), ha attribuito un peso maggiore alla classe minoritaria (partecipanti) durante l'apprendimento.
La scelta di XGBoost è motivata dalla sua capacità di gestire relazioni non lineari e interdipendenze complesse in dataset strutturati di medie dimensioni, come quello in esame, grazie alla sua efficienza computazionale e alla robustezza al rumore, caratteristiche validate dagli autori dell'algoritmo (Chen & Guestrin, 2016). Questa metodologia si adatta bene all'analisi di dati agricoli, come quelli relativi a variabili strutturali e geografiche (età del titolare, SAU, OTE, ecc.), grazie alla sua capacità di ottimizzare sequenze di alberi decisionali tramite gradient boosting, minimizzando la funzione di perdita in modo iterativo.
Tuttavia, indipendentemente dalla qualità dell'algoritmo utilizzato, uno dei principali limiti degli algoritmi di machine learning è la scarsa interpretabilità dei risultati. Questo significa che spesso i modelli sono percepiti come "scatole nere", in cui le relazioni tra le variabili e le previsioni non sono facilmente comprensibili (Rudin, 2019).
Sebbene XGBoost fornisca metriche come l'importanza delle variabili, queste offrono solo una visione complessiva del loro contributo: non permettono, cioé di capire con precisione come ogni variabile influisca sulle singole previsioni, né indicano se l'effetto di una variabile sia positivo o negativo.
Questo è particolarmente problematico in contesti applicativi come questo, dove non è sufficiente fare previsioni ma è fondamentale comprendere e giustificare il ruolo di ciascuna variabile per supportare le decisioni politiche e informare i policy maker.
Per affrontare questo problema si è optato per l'impiego di SHAP (SHapley Additive exPlanations), un metodo che si basa sulla teoria dei giochi cooperativi.
SHAP attribuisce un valore specifico a ciascuna variabile per ogni previsione, permettendo di interpretare i risultati del modello XGBoost in modo chiaro e trasparente. In altre parole, il valore

[7] L'overfitting si verifica quando un modello impara troppo bene i dettagli e i rumori presenti nei dati su cui è stato addestrato, ma poi fatica a fare previsioni corrette su nuovi dati. In pratica, il modello diventa troppo complesso e si adatta troppo strettamente ai dati di addestramento, perdendo la capacità di generalizzare a situazioni reali. Di conseguenza, funziona bene sui dati di addestramento, ma meno bene quando lo si testa su nuovi dati.

SHAP è una misura di distribuzione equa delle "ricompense" o "punteggi" in un gioco cooperativo, che assegna a ciascun partecipante (in questo caso, una variabile del modello) il suo contributo medio in tutte le possibili combinazioni di partecipanti. Nel contesto di SHAP, il gioco è il modello predittivo e i "partecipanti" sono le singole variabili del modello. Il valore di Shapley per una variabile j, denotato come φj , si calcola come la media ponderata del cambiamento nelle previsioni del modello quando si aggiunge la variabile j a tutte le possibili combinazioni delle altre.

La formula per il calcolo del valore di Shapley di un feature j è la seguente:

$$\phi j(f) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)]$$

dove (N) è l'insieme di tutte le variabili, (S) è un sottoinsieme di variabili esclusa (j), e (f(S)) è la previsione del modello considerando solo le variabili in (S).
Il vantaggio di SHAP risiede nella sua capacità di offrire interpretazioni sia globali, aggregando i contributi su tutto il dataset, sia locali, analizzando singole previsioni, rendendolo superiore ad altri metodi come LIME[8], valido per singole previsioni, o il feature importance nativo di XGBoost, che calcola solo il valore medio complessivo (Lundberg e Lee, 2017).
Dopo aver calcolato i valori SHAP medi per tutte le variabili, è stata determinata la baseline SHAP, ovvero il valore medio delle previsioni del modello, che sarà utilizzato per l'interpretazione dei risultati. Inoltre, è stata costruita una tabella di confronto che mette in relazione i valori SHAP e la corrispondente probabilità di partecipazione, utile per l'analisi dei risultati.

[8] LIME (Local Interpretable Model-Agnostic Explanations) è un metodo di interpretabilità che si focalizza sull'analisi di singoli casi. Per ogni osservazione, LIME crea un modello lineare locale (un "surrogato") che approssima il comportamento del modello originale in un intorno dell'osservazione. L'interpretazione si basa sull'analisi dei coefficienti di questo modello locale, che indicano l'importanza relativa delle features per la previsione di quel caso specifico.

# RISULTATI DELL'ANALISI

## VALUTAZIONE DEL MODELLO E INTERPRETAZIONE DELLE METRICHE

La valutazione del modello è un passaggio essenziale per comprenderne l'accuratezza e l'affidabilità. Le metriche qui calcolate includono l'AUC-ROC (Area Under the Receiver Operating Characteristic Curve) e l'F1-score, che unitamente all'analisi della Log Loss, forniscono informazioni complementari sulla capacità del modello di distinguere la partecipazione delle aziende agricole ai bandi.

L'AUC-ROC ottenuto è pari a 0,8192, indicando una buona capacità discriminante. Questo parametro misura la capacità del modello di assegnare punteggi più alti alle aziende che partecipano rispetto a quelle che non partecipano. Un valore di 0,5 equivarrebbe a una scelta casuale, mentre un valore di 1 indicherebbe una separazione perfetta. Il risultato ottenuto suggerisce che il modello distingue correttamente la maggior parte delle aziende partecipanti da quelle non partecipanti, coerentemente con la letteratura che associa valori superiori a 0,8 a una buona discriminazione (Fawcett, 2006).

L'F1-score calcolato, invece, è pari a 0,8134, un valore che conferma un buon equilibrio tra *precision*, ovvero la proporzione di aziende classificate come partecipanti che effettivamente hanno partecipato e *recall*, cioè la proporzione di aziende partecipanti correttamente identificate (Powers, 2020). Questo equilibrio è particolarmente importante in un dataset sbilanciato come quello in esame, dove le aziende non partecipanti sono molto più numerose.

Infine, la Log Loss ottenuta è di 0,5144 sul set di test e 0,4981 sul set di addestramento. La Log Loss è una misura dell'errore del modello nel prevedere le probabilità associate alla classe corretta, penalizzando maggiormente le previsioni molto errate (Reid e Williamson, 2011); la differenza non troppo elevata tra i valori dei due set suggerisce una buona calibrazione senza overfitting.

A fronte di questi risultati, si può affermare che il modello è in grado di prevedere la probabilità di partecipazione ai bandi con una buona accuratezza, garantendo un bilanciamento efficace tra sensibilità e precisione. La sua affidabilità è ulteriormente supportata dal corretto trattamento della classe di minoranza e dall'analisi della Log Loss, che assicura una buona generalizzabilità.

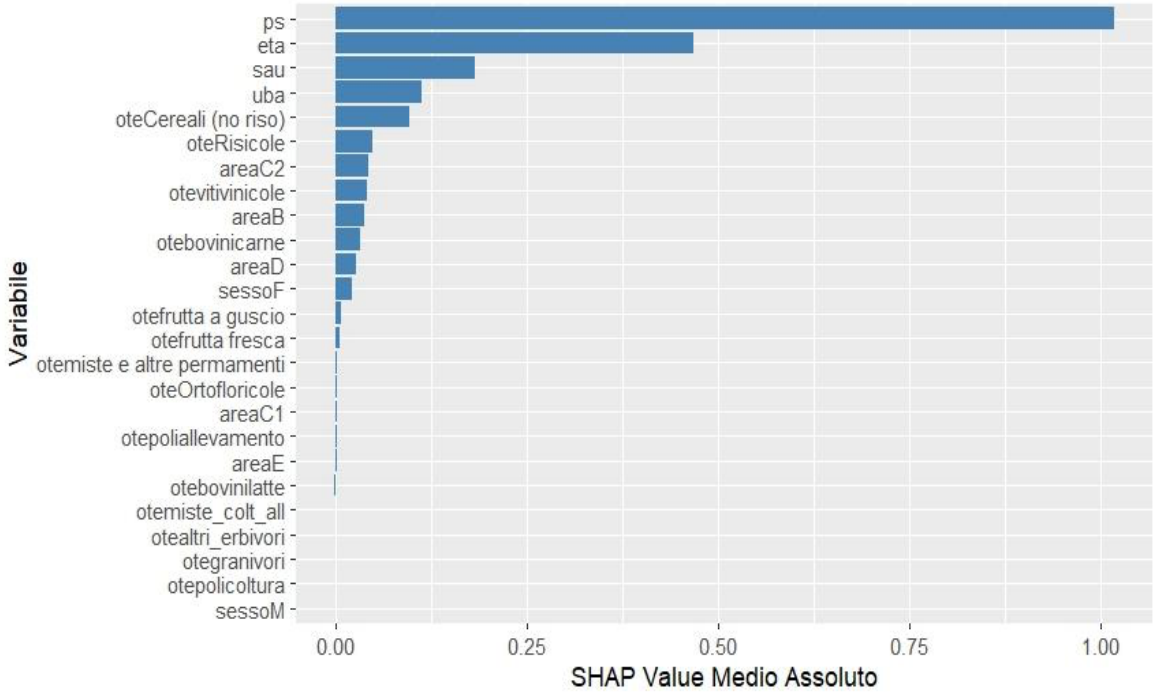
## IMPORTANZA MEDIA ASSOLUTA DELLE VARIABILI (FEATURE) CON SHAP

L'analisi dell'importanza delle variabili, basata sui valori SHAP (fig.1) permette di capire quanto ogni caratteristica contribuisca, in media, a distinguere le aziende agricole che partecipano al bando pubblico da quelle che non partecipano. L'importanza media assoluta di una variabile, calcolata come la media dei valori SHAP assoluti su tutte le aziende, misura quanto quella caratteristica sia rilevante per il modello nel separare i due gruppi (partecipanti e non partecipanti), indipendentemente dalla direzione dell'effetto. Una variabile con un valore SHAP medio assoluto alto è quindi significativa per il modello, ma non implica necessariamente che favorisca la partecipazione: potrebbe essere altrettanto importante per identificare chi non partecipa, e questo è un aspetto rilevante per comprendere le dinamiche dei bandi sull'Operazione 4.1.1 oggetto di analisi.

Ad esempio, tra gli OTE, quello che riguarda le aziende specializzate in cerali ("oteCereali (no riso)") ha un valore SHAP medio assoluto di 0,097, mentre l'OTE delle aziende vitivinicole ("otevitivinicole") ha un valore di 0,041, suggerendo che le aziende specializzate nella coltivazione dei cerali ricoprano un ruolo più rilevante per il modello nel distinguere i due gruppi.

Osservando i dati reali sull'incidenza dei gruppi di aziende per OTE rispetto alla partecipazione e alla non partecipazione (tab. 2), emerge che le aziende cerealicole, pur rappresentando il 15% delle aziende piemontesi, incidano solo per il 5% sul totale dei partecipanti ai bandi sulla 4.1.1. Al contrario, le aziende specializzate in viticoltura da vino, rappresentando il 18% delle aziende regionali, incidano addirittura per il 23% sul totale delle aziende partecipanti.

**Figura 1 - Valori SHAP medi assoluti per le variabili utilizzate nel modello xgboost**




Fonte: Elaborazione IRES Piemonte

**Tabella 2 - Composizione percentuale delle aziende agricole piemontesi per OTE e incidenza percentuale dei partecipanti e non partecipanti ai bandi PSR 4.1.1.**

| OTE | Incidenza aziende sul totale | Incidenza non partecipanti | Incidenza partecipanti |
|---|---|---|---|
| *Vitivinicole* | *17,94%* | *17,46%* | *22,70%* |
| Bovini da carne | 9,29% | 9,04% | 11,76% |
| Altri seminativi | 16,54% | 17,39% | 8,18% |
| Bovini da latte | 2,42% | 1,88% | 7,73% |
| Frutta a guscio | 8,52% | 8,68% | 6,90% |
| Frutta fresca | 3,92% | 3,62% | 6,88% |
| Miste coltiv. allev. | 5,63% | 5,57% | 6,27% |
| Policoltura | 6,22% | 6,29% | 5,47% |
| *Cereali (no riso)* | *14,74%* | *15,73%* | *4,97%* |
| Miste e altre permamenti | 4,59% | 4,59% | 4,62% |
| Altri erbivori | 3,09% | 2,98% | 4,12% |
| Granivori | 1,03% | 0,75% | 3,80% |
| Ortofloricole | 2,67% | 2,60% | 3,37% |
| Risicole | 3,29% | 3,34% | 2,79% |
| Poliallevamento | 0,12% | 0,08% | 0,45% |
| **Totale** | **100,00%** | **100,00%** | **100,00%** |

Fonte: Elaborazione IRES su dati Anagrafe Agricola Unica del Piemonte e CSI Sistema di Monitoraggio PSR

Questo indica che, nonostante il valore SHAP medio assoluto più alto, "oteCereali (no riso)" è più significativa per identificare le aziende che non partecipano, piuttosto che quelle che partecipano. In altre parole, questa variabile è utile al modello per classificare i casi di non partecipazione. Le aziende vitivinicole, dal canto loro, pur avendo un valore SHAP medio assoluto

più basso, mostrano una partecipazione reale più alta, ma la loro capacità di discriminazione è in parte "assorbita" da altre variabili più influenti, come la produzione standard (PS), che ha il valore SHAP medio assoluto più alto (1,018). Per comprendere se una variabile significativa favorisca o penalizzi la partecipazione, è quindi necessario utilizzare degli strumenti specifici quali, ad esempio i dependence plot.

# IL RUOLO INTERPRETATIVO DEI DEPENDENCE PLOT

I dependence plot sono strumenti grafici fondamentali per visualizzare la relazione tra una variabile indipendente e il suo impatto sulle previsioni del modello attraverso i valori SHAP. Ogni punto nel grafico rappresenta un'osservazione del dataset, posizionata in base al valore della variabile di interesse sull'asse orizzontale e al valore SHAP corrispondente sull'asse verticale. Un valore SHAP positivo indica che la variabile aumenta la probabilità di partecipazione rispetto alla media, mentre un valore SHAP negativo indica che la riduce.
Questi grafici sono essenziali per capire, quindi, se una variabile identificata dall'importanza media assoluta come significativa per il modello, favorisca o penalizzi la partecipazione e per analizzare come l'effetto di una variabile possa variare in combinazione con altre, evidenziando interazioni e non linearità nelle relazioni.
L'interpretazione dei dependence plot varia a seconda che la variabile analizzata sia continua o categorica. Per le variabili continue, il grafico consente di individuare andamenti globali, come relazioni lineari, effetti soglia o cambiamenti di tendenza.
Per le variabili categoriche, invece, il dependence plot assume un aspetto di tipo box-plot, con ciascuna categoria rappresentata su una posizione distinta dell'asse orizzontale, permettendo di confrontare l'impatto relativo delle diverse categorie.
Per interpretare la dependence plot, è fondamentale comprendere come i valori SHAP si traducano in probabilità di partecipazione. Il modello, infatti, produce una previsione in termini di rapporto logaritmico tra la probabilità di partecipazione e quella di non partecipazione (log-odds) e il valore necessita di essere trasformato in probabilità tramite la formula:

$$probabilità = \frac{e^{log-odds}}{1 + e^{log-odds}}$$

In assenza di informazioni specifiche sulle variabili, il modello assegna un valore medio di log-odds, chiamato baseline logit, che in questo caso è pari a -0,87 (valore SHAP). Operando la trasformazione risulta, infine, una probabilità media di partecipazione del 29,5% che rappresenta la probabilità di partecipazione al bando media di un'azienda piemontese senza che tale probabilità sia modificata da alcuna delle caratteristiche inserite nel modello.

Per semplificare la lettura dei plot, la baseline logit (-0,87) è stata riparametrata al valore zero sull'asse y, in modo che i valori SHAP rappresentino una variazione positiva o negativa rispetto a questa baseline. La tabella 3 riporta la corrispondenza tra valori SHAP, probabilità di partecipazione e distanza in punti percentuali dalla baseline, offrendo una guida pratica per interpretare i grafici delle sezioni successive.

**Tabella 3 - Valori SHAP, probabilità e differenza dalla baseline in punti percentuale**

| Valore SHAP | Probabilità (%) | Differenza dalla baseline (pp) |
|---|---|---|
| -2 | 5,40% | -24,1 |
| -1,8 | 6,50% | -23 |
| -1,6 | 7,80% | -21,7 |
| -1,4 | 9,40% | -20,1 |
| -1,2 | 11,20% | -18,3 |
| -1 | 13,40% | -16,1 |
| -0,8 | 15,90% | -13,6 |
| -0,6 | 18,70% | -10,8 |
| -0,4 | 21,90% | -7,6 |
| -0,2 | 25,50% | -4 |
| **0,0** | **29,50%** | **0** |
| 0,2 | 33,90% | 4,4 |
| 0,4 | 38,50% | 9 |
| 0,6 | 43,30% | 13,8 |
| 0,8 | 48,30% | 18,8 |
| 1 | 53,30% | 23,8 |
| 1,2 | 58,20% | 28,7 |
| 1,4 | 63,00% | 33,5 |
| 1,6 | 67,50% | 38 |
| 1,8 | 71,70% | 42,2 |
| 2 | 75,60% | 46,1 |

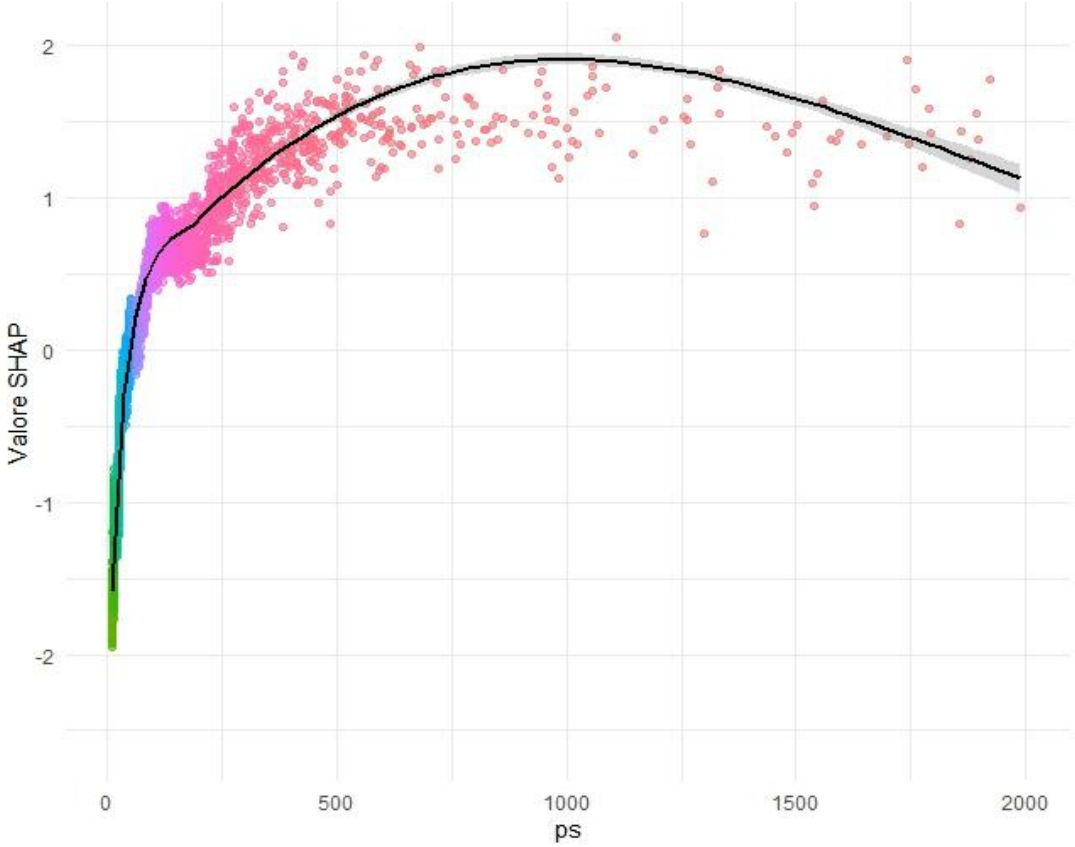
Fonte: Elaborazione IRES Piemonte

Nei paragrafi successivi, infatti, saranno analizzati in dettaglio i dependence plot delle variabili più rilevanti per il modello, ordinate per importanza in base al valore SHAP medio assoluto (cfr. fig. 1), mostrando come ciascuna caratteristica influisca sulla probabilità di partecipazione. Inoltre, saranno esplorate le interazioni tra variabili, in particolare l'effetto combinato di "ote Cereali (no riso)" ed "ote vitivinicole" con la produzione standard (PS) per comprendere meglio le dinamiche che influenzano la partecipazione al bando e fornire un esempio della grande flessibilità di questi metodi analitici.

## EFFETTO DELLA PRODUZIONE STANDARD

La produzione standard (PS), con un valore SHAP medio assoluto di 1,018, è la variabile più importante per il modello. Il dependence plot (Fig. 2) mostra come il valore SHAP vari al crescere della produzione standard da 0 a 2 milioni di euro. Per le aziende con valori di PS fino a 70 mila euro, il valore SHAP è fortemente negativo (fino a -2, con una media pari a -1,060), riducendo significativamente la probabilità di partecipazione. Man mano che la PS aumenta, il valore SHAP cresce rapidamente, diventando stabilmente positivo intorno al valore di 72 mila euro. La probabilità di partecipazione raggiunge il suo massimo per le aziende con una PS di circa 1,1 milioni di euro, dando a queste una probabilità di partecipazione superiore al 75%. Oltre questa soglia, pur rimanendo sopra la baseline, i valori SHAP mostrano una flessione. Questa dinamica può essere spiegata dal fatto che le aziende di modeste dimensioni economiche non sono propense a partecipare perché non hanno sufficiente capitale circolante necessario al cofinanziamento e/o non riescono ad accedere o a sostenere un credito, a differenza di quelle di dimensioni medio-grandi che mostrano probabilità di partecipazione molto elevate. Infine, le aziende capitalistiche, pur avendo alte probabilità di partecipazione, possono essere meno attratte rispetto a quelle medio-grandi, forse per i limiti di spesa previsti dai bandi.

**Figura 2 Dependence plot relativo alla produzione standard aziendale**

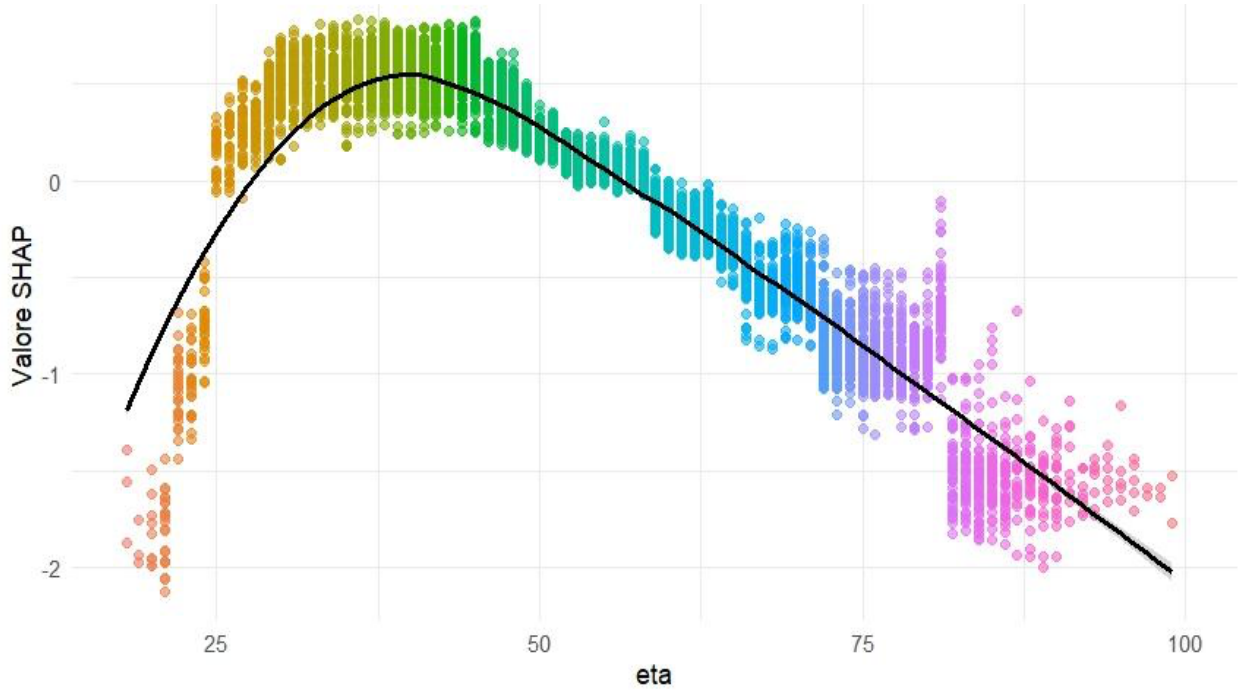



Fonte:-Elaborazione IRES Piemonte

## EFFETTO DELL'ETÀ DELL'IMPRENDITORE AGRICOLO

L'età dell'agricoltore (fig. 3), con un valore SHAP medio assoluto di 0,468, è la seconda variabile più importante per il modello. Per gli agricoltori molto giovani (20-30 anni), il valore SHAP è positivo (fino a +1; 54% di probabilità), indicando un aumento della probabilità di partecipazione. Tuttavia la probabilità di partecipazione diminuisce progressivamente con l'aumentare dell'età, e il valore SHAP per età superiori di 50 anni diventa negativo, indicando una probabilità di partecipazione sotto la media. Per le aziende il cui titolare ha superato gli 80 anni, il valore SHAP raggiunge un minimo di circa -2, che significa una probabilità di partecipazione pressoché azzerata. Il plot per questa variabile dimostra chiaramente che gli agricoltori più giovani sono più propensi a partecipare a questo tipo di bando, mentre quelli più anziani tendono a partecipare meno, con un effetto particolarmente marcato oltre i 50 anni.

**Figura 3 - Dependence plot relativo all'età dell'imprenditore agricolo**

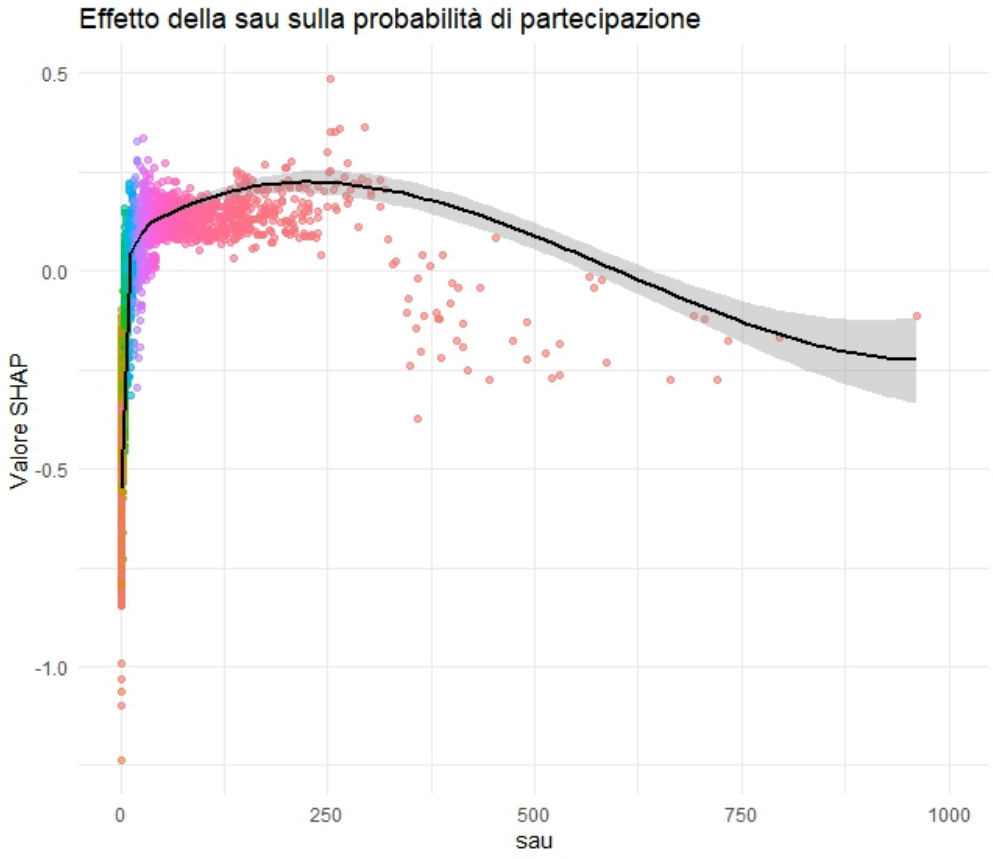



Fonte:-Elaborazione IRES Piemonte

## EFFETTO DELLA SUPERFICIE AGRICOLA UTILIZZATA (SAU)

La Superficie Agricola Utilizzata, presenta un valore SHAP medio assoluto di 0,182. Il dependence plot (Fig. 4) mostra come il valore SHAP vari al crescere della superficie, da 0 a 500 ettari.
Per valori di "sau" molto bassi (fino a circa 10 ettari), il valore SHAP è negativo, riducendo la probabilità di partecipazione. Man mano che la superficie coltivata dall'azienda aumenta, il valore SHAP cresce rapidamente diventando positivo per valori superiori e raggiungendo un massimo di circa +0,5, equivalente a una probabilità di partecipazione del 50%, quando la SAU raggiunge i 250 ettari. Al di sopra di tali dimensioni, comunque non frequenti in Piemonte dove la SAU aziendale media si attesta intorno ai 20 ettari e la presenza di piccole realtà è ancora molto forte, la probabilità inizia a diminuire e torna a essere negativa a partire da dimensioni di 600 ettari.
Analizzando la scala di valori SHAP per la SAU, risulta in ogni caso che tale variabile non incide fortemente sulle probabilità di partecipare considerato che il valore SHAP nella sua parte positiva oscilla tra 0 e 0,48, cioè tra il 29,5% della baseline e il 40% del punto di massima probabilità.

**Figura 4 - Dependence plot relativo alla SAU aziendale**



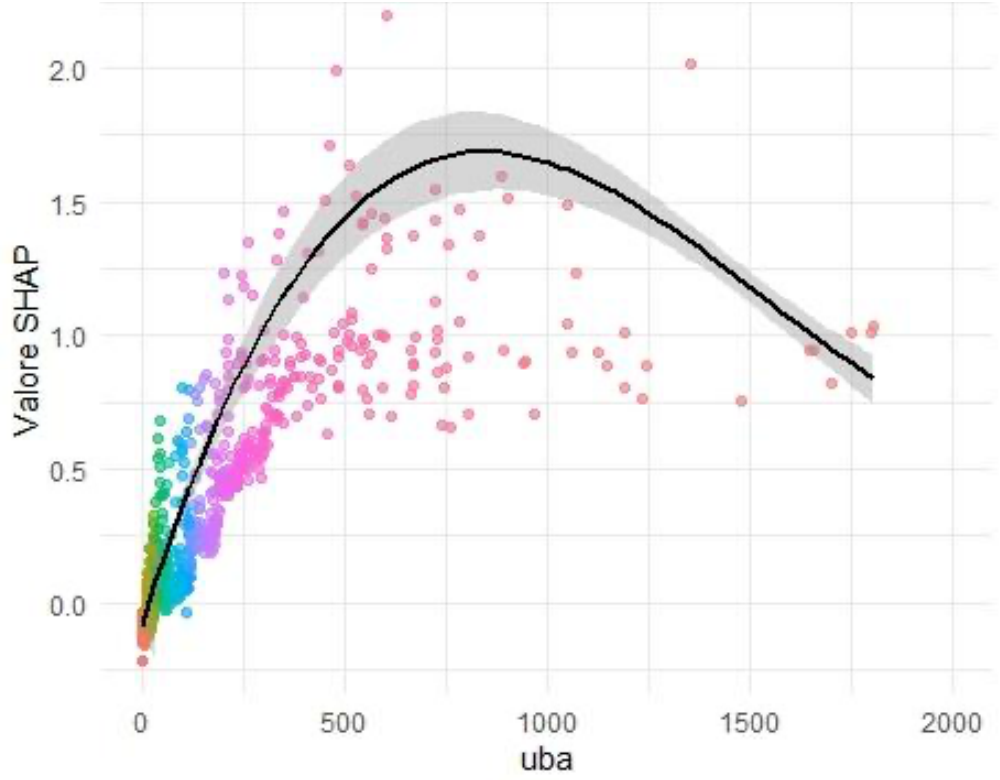

Fonte:-Elaborazione IRES Piemonte

## EFFETTO DELLE UNITÀ DI BESTIAME ADULTO (UBA)

Anche l'aumento della consistenza degli allevamenti, misurata tramite il valore delle UBA, fa aumentare la probabilità di partecipazione ai bandi sull'Operazione 4.1.1, e come già illustrato nell'analisi delle precedenti variabili la relazione non segue un andamento lineare.

Partendo da un valore SHAP medio assoluto pari a 0,113 (probabilità = 32%), Il dependence plot (Fig. 5) evidenzia un valore negativo solo per gli allevamenti di dimensioni molto piccole che delineano profili di aziende che difficilmente possono avere spazio di mercato, ma per le restanti aziende anche modeste il valore SHAP aumenta rapidamente fino a raggiungere un picco di circa +1,5 per le aziende intorno alle 1000 UBA, il che significa che queste hanno una probabilità di partecipazione del 65% circa: 36 punti percentuale in più della baseline.

**Figura 5 - Dependence plot relativo alla dimensione degli allevamenti**




Fonte:-Elaborazione IRES Piemonte

## EFFETTO DEGLI ORIENTAMENTI TECNICO ECONOMICI

Le variabili legate all'orientamento tecnico economico (OTE) mostrano impatti differenziati sulla probabilità di partecipazione ai bandi, come evidenziato dai dependence plot riportati in allegato I.

Le aziende cerealicole (oteCereali, no riso), con un valore SHAP medio assoluto di 0,097, sono tra le più influenti, ma come detto, questa importanza è utile al modello per discriminare i non partecipanti. Il dependence plot indica, infatti, che le aziende cerealicole (valore 1) hanno un valore SHAP medio di -0,428 che implica una probabilità di partecipazione al 19% (-10,5 punti percentuali rispetto alla baseline), mentre le altre aziende (valore 0) hanno un valore SHAP medio di 0,039, aumentando la probabilità al 30,5% (+1 punto percentuale). Analogamente, le aziende risicole (oteRisicole), con un valore SHAP medio assoluto simile, mostrano un effetto ancora più penalizzante: il dependence plot evidenzia un valore SHAP medio di -0,67 che si traduce in una probabilità di partecipazione del 16% (-13,5 punti percentuali rispetto alla baseline). Anche le aziende di bovini da carne (oteBoviniCarne), con un valore SHAP medio assoluto di 0,032, presentano un impatto negativo: il dependence plot presenta, infatti, un valore SHAP medio di -0,141 che assegna una probabilità di partecipazione del 26% (-3,5 punti percentuali). Le aziende di bovini da latte (oteBoviniLatte), invece hanno un impatto quasi nullo rispetto alla baseline. Il relativo dependence plot indica un valore SHAP medio di 0,007, ovvero una probabilità di partecipazione del 30%. Impatti positivi, invece, si evidenziano per tutto il gruppo di di coltivazioni permanenti.

Le aziende di frutta fresca mostrano un effetto positivo, con una probabilità di partecipazione del 31% e un valore del tutto analogo a quello delle aziende di frutta a guscio. Infine, per le aziende vitivinicole si registra una probabilità di partecipazione media del 32%.

Questi risultati, supportati da quelli in tabella 2, confermano che, per la maggior parte degli OTE la probabilità di partecipazione non si discosta molto dalla baseline (circa 29,5%-30%), ma che i comparti più penalizzati dal punto di vista della probabilità di partecipazione sono quelli orientati alle produzioni commodity, mentre quelli che possono più facilmente produrre prodotti di qualità certificate e a maggior valore aggiunto hanno, al contrario, probabilità maggiori.

## EFFETTO DELLA LOCALIZZAZIONE AZIENDALE

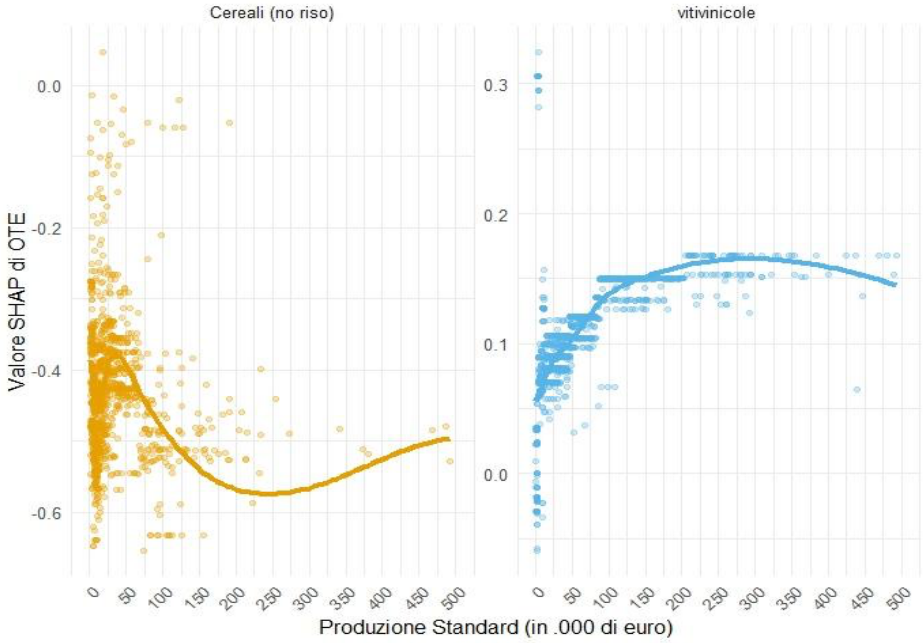
La localizzazione aziendale, rappresentata dalle aree geografiche del Piemonte, non mostra un impatto significativo sulla probabilità di partecipazione, ma dai dependence plot in Allegato 2, si può osservare come le aziende situate in aree di alta collina e montane, tendono ad avere maggiori probabilità di partecipazione rispetto a quelle in aree di pianura.
Nello specifico l'area C2 presenta un valore SHAP medio di 0,211 che corrisponde a una probabilità di 34,1%: 4,6 punti percentuale al di sopra della baseline, anche le aziende in area D mostrano una probabilità stimata superiore alla baseline, 32,1%, più contenuta delle aree C2.
Nelle aree B, invece il valore SHAP medio negativo, indica una probabilità inferiore alla baseline, ma comunque davvero lieve: 28,3%. Nelle aree C1, infine, il valore medio SHAP è in linea con la baseline.
Questo risultato potrebbe riflettere non solo le differenze settoriali, ma anche fattori legati al fatto che in aree montane gli agricoltori sono mediamente più giovani. Inoltre può essere stato determinante il fatto che le aziende di questi territori godano di un cofinanziamento maggiore per l'investimento, alla stregua di ciò che accade per i giovani.

# L'EFFETTO SINERGICO DI PIÙ VARIABILI

I plot presi in considerazione fino ad ora evidenziano l'effetto sulla probabilità di partecipazione focalizzando l'attenzione su variabili prese singolarmente. Tuttavia, è molto interessante mettere in evidenza gli effetti che una variabile esercita su di un'altra.
In questa sezione si offre un approfondimento sul tema, anche con finalità esemplificative rispetto la grande flessibilità del modello, analizzando come l'effetto degli OTE qui più discussi- ceralicereali e vitivincole-vitivinicole cambino al variare del parametro della produzione standard.
Come evidenziato, la produzione standard è la variabile più importante per il modello, ma il suo effetto non è uguale per tutte le attività agricole. Per le aziende che coltivano cereali, il grafico a sinistra rivela che questa attività riduce costantemente la probabilità di partecipazione, indipendentemente dal suo valore, rimanendo costantemente al di sotto della linea di baseline, ma per le aziende con maggiore produzione standard la curva sale aumentando le probabilità delle aziende rispetto a quelle di dimensioni economiche più piccole.
Per le aziende vitivinicole (grafico a destra), invece, l'effetto positivo emerge già per aziende con valori di produzione standard moderate, mostrando però, contrariamente alle aziende cerealicole, una flessione per i livelli di produzione standard maggiori.
Questi risultati, confermando che il modello considera le interazioni tra variabili e la non linearità delle relazioni, suggeriscono ai policy makers che i bandi hanno attirato soprattutto aziende

con una produzione elevata, come quelle vitivinicole, ma sono risultati meno appetibili per chi coltiva cereali, anche in aziende più strutturate. Si potrebbe quindi pensare a incentivi specifici per le aziende con cereali, ad esempio semplificando i requisiti per chi ha una produzione standard più bassa o offrendo supporto per aumentare il loro valore economico.

**Figura 6 - Relazione tra il valore SHAP e la produzione standard per due categorie di attività agricole: cereali (senza riso) a sinistra e vitivinicole a destra.**




Fonte: Elaborazione IRES Piemonte

# CONCLUSIONI E RACCOMANDAZIONI

## CONCLUSIONI

Questa ricerca ha esplorato i fattori che influenzano la partecipazione delle aziende agricole piemontesi ai bandi dell'Operazione 4.1.1 del PSR 2014-2022, adottando un approccio basato sul Machine Learning (ML) con l'algoritmo XGBoost, integrato dall'interpretazione tramite i valori SHAP.

L'analisi ha permesso di identificare come variabili economiche, strutturali e, in misura minore, geografiche giochino un ruolo cruciale nel determinare la probabilità di partecipazione, offrendo una base solida per comprendere le dinamiche sottostanti.

Tra queste, la produzione standard (PS), espressa in euro nel dataset originale, emerge come il fattore più influente, caratterizzato da un effetto non lineare che varia significativamente in base alla dimensione economica dell'azienda.

In particolare, le aziende con una PS inferiore a 70.000 euro mostrano una probabilità di partecipazione notevolmente ridotta, come evidenziato dai valori SHAP fortemente negativi (fino a -2), il che suggerisce una barriera significativa per le realtà di dimensioni più modeste.

Al contrario, le aziende di medie e grandi dimensioni, con una PS che raggiunge circa 1,1 milioni di euro, presentano una probabilità di partecipazione che supera il 75%, indicando un picco di attrattività per questo tipo di bandi.

Tuttavia, oltre questa soglia, i valori SHAP mostrano una flessione, pur rimanendo sopra la baseline del 29,5%, il che potrebbe riflettere una minore propensione delle aziende di dimensioni molto elevate, forse a causa di vincoli specifici dei bandi, come limiti di spesa, anche se questa interpretazione rimane ipotetica senza dati aggiuntivi.

Un altro elemento chiave emerso dall'analisi è l'età dell'imprenditore agricolo, che influenza in modo marcato la probabilità di partecipazione. Gli agricoltori più giovani, tra i 20 e i 30 anni, registrano valori SHAP positivi (fino a +1), corrispondenti a una probabilità del 54%, suggerendo una maggiore inclinazione a cogliere opportunità offerte dai bandi.

Al contrario, la probabilità diminuisce progressivamente con l'aumentare dell'età, diventando negativa oltre i 50 anni e raggiungendo un minimo di circa -2 per gli ultraottantenni, con una probabilità di partecipazione quasi nulla.

Questo trend potrebbe indicare che i giovani sono più aperti a innovazioni e processi amministrativi, mentre gli agricoltori più anziani potrebbero affrontare ostacoli legati alla familiarità con le procedure o alla propensione al cambiamento, sebbene tali considerazioni restino speculative senza evidenze dirette.

Inoltre, l'età avanzata potrebbe preludere a una chiusura aziendale o a un subentro generazionale, un aspetto che merita ulteriori indagini per confermare se il subentro possa effettivamente rivitalizzare la partecipazione.

L'analisi delle variabili legate all'orientamento tecnico economico (OTE) e alla localizzazione aziendale, approfondita attraverso i dependence plot, ha messo in luce alcune disparità settoriali e territoriali. Le aziende cerealicole e risicole presentano una bassa probabilità di partecipazione, con valori SHAP medi negativi (rispettivamente -0,428 e -0,67), che si traducono in probabilità del 19% e 16%, riflettendo una tendenza a essere escluse dai bandi.

Al contrario, le aziende vitivinicole mostrano una probabilità più alta, intorno al 32%, specialmente quando associate a una PS elevata (600.000 euro).

Per quanto riguarda la localizzazione, le aziende situate in aree collinari e montane (aree C2 e D) registrano probabilità superiori alla baseline (rispettivamente 34,1% e 32,1%), mentre quelle in pianura (area B) sono leggermente sotto (28,3%). Tuttavia, la localizzazione non emerge come una variabile dominante nel modello, indicando che il suo impatto è secondario rispetto a fattori economici e strutturali.

In sintesi, i risultati suggeriscono che i bandi dell'Operazione 4.1.1 attraggono soprattutto aziende di medie e grandi dimensioni operanti in comparti a maggior valore aggiunto, come la viticoltura, mentre penalizzano quelle cerealicole e risicole, spesso legate a produzioni di commodities. Questa dinamica evidenzia una possibile selezione naturale verso realtà più strutturate, ma solleva interrogativi sulle opportunità per i comparti meno rappresentati.

Un aspetto che richiede ulteriori approfondimenti è il ruolo dei criteri di selezione utilizzati per assegnare i punteggi e formare le graduatorie dei bandi. Variabili come la PS, la superficie agricola utilizzata (SAU), l'età del titolare e l'OTE sono frequentemente incluse nei sistemi di punteggio, e un'analisi futura basata su un approccio ML simile a quello adottato in questo studio potrebbe rivelare se tali criteri favoriscano o penalizzino eccessivamente determinati settori o caratteristiche aziendali, fornendo indicazioni utili per una loro eventuale revisione e un miglioramento della distribuzione delle risorse.

# RACCOMANDAZIONI

Sulla base dei risultati emersi dall'analisi, si propongono una serie di raccomandazioni che potrebbero essere prese in considerazione dai programmatori e dai responsabili delle politiche agricole per ottimizzare la partecipazione ai bandi del PSR e affrontare le disuguaglianze settoriali e strutturali osservate. Queste raccomandazioni mirano a promuovere un accesso più equo e a massimizzare l'impatto degli incentivi pubblici, bilanciando le esigenze di diverse tipologie di aziende agricole.

- **Incentivi mirati per le aziende con minori probabilità di partecipazione.** Le aziende con una produzione standard (PS) inferiore a 70.000 euro sono fortemente svantaggiate, come indicato dai valori SHAP negativi fino a -2, che riducono la loro probabilità di partecipazione a livelli molto bassi. Per supportare queste realtà, si potrebbe valutare l'introduzione di un cofinanziamento più favorevole, che allevi il carico finanziario iniziale. Inoltre, l'organizzazione di programmi di formazione potrebbe aiutare questi agricoltori a gestire meglio le pratiche burocratiche e a comprendere le opportunità offerte dai bandi, stimolando così la loro partecipazione. Un'ulteriore misura potrebbe consistere nella semplificazione dei requisiti di accesso per chi ha una PS più bassa, ad esempio riducendo la documentazione richiesta, o nell'offerta di supporto finanziario tramite agevolazioni per l'accesso al credito, che potrebbero compensare la mancanza di capitale circolante. Queste azioni potrebbero favorire l'inclusione di aziende più piccole, che altrimenti rimangono escluse a causa di barriere strutturali.

- **Promozione della partecipazione tra gli agricoltori di età matura.** L'analisi evidenzia che la base imprenditoriale agricola piemontese è ancora largamente composta da agricoltori in età matura, con una probabilità di partecipazione che declina significativamente oltre i 50 anni, fino a quasi azzerarsi oltre gli 80 anni, come mostrato dai valori SHAP negativi fino a -2. Pur riconoscendo che la maggior parte degli agricoltori si affida ai CAA gestiti dalle organizzazioni di categoria per la compilazione delle domande, si potrebbero esplorare opportunità specifiche per incentivare la partecipazione degli anziani. Una soluzione potrebbe consistere nello sviluppo di maschere di compilazione dei bandi semplificate, progettate per essere intuitive e accessibili anche per chi ha meno familiarità con le procedure digitali o amministrative, riducendo così le barriere all'ingresso. Altrettanto utili potrebbero essere strumenti innovativi, come guide interattive o piattaforme digitali con supporto guidato, che facilitino la presentazione delle domande in modo autonomo o con un supporto minimo. Inoltre, una forte animazione territoriale sulle opportunità offerte dai bandi, attraverso incontri informativi e campagne di sensibilizzazione mirate, potrebbe aumentare la consapevolezza degli agricoltori più anziani sui benefici potenziali, incoraggiandoli a coinvolgere i CAA o a partecipare direttamente. Questo approccio potrebbe non solo stimolare la loro partecipazione, valorizzando la loro profonda conoscenza del territorio e delle pratiche tradizionali, ma anche favorire un dialogo intergenerazionale, contribuendo a mitigare il rischio di chiusura aziendale e a trasmettere competenze alle nuove generazioni.

- **Riorientamento degli incentivi per evitare il deadweight.** L'analisi dei dependence plot per la produzione standard (PS), la superficie agricola utilizzata (SAU) e le unità di bestiame adulto (UBA) rivela che le aziende con una PS intorno a 1,1 milioni di euro, una

SAU superiore a 250 ettari (punto di picco) e un numero di UBA intorno a 1000 presentano probabilità di partecipazione molto elevate (fino al 75% e 65%, rispettivamente). Queste realtà, grazie alla loro dimensione economica e strutturale, potrebbero già disporre delle risorse necessarie per investire senza dipendere dal cofinanziamento pubblico, sollevando il rischio di deadweight, ovvero l'assegnazione di fondi a soggetti che avrebbero comunque effettuato gli investimenti. Per massimizzare l'efficacia degli incentivi, si potrebbe considerare una rimodulazione delle risorse, concentrandole su aziende con dimensioni più modeste, inferiori ai valori di picco. Queste aziende, pur avendo una probabilità di partecipazione più bassa, potrebbero trarre un beneficio significativo dal supporto pubblico.

# BIBLIOGRAFIA

Adamo, M., Cavaletto, S. (2024). Piemonte rurale 2024. IRES Piemonte.

Adamo, M., Torchio, N. (2025). Il ricambio generazionale nel PSR 2014–2022 del Piemonte. Note sullo sviluppo rurale, IRES Piemonte.

Araújo, S. O., Peres, R. S., Barata, J., Lidon, F., & Ramalho, J. C. (2023). Machine learning applications in agriculture: Current trends, challenges, and future perspectives. Agronomy, 13(2), 459. https://doi.org/10.3390/agronomy13020459

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 248–266. https://doi.org/10.1145/3531146.3533090

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 69–78. https://doi.org/10.1145/1014052.1014063

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Chy, Md. K. H., & Buadi, O. N. (2024). Role of machine learning in policy making and evaluation. International Journal of Innovative Science and Research Technology, 9(10). https://doi.org/10.38124/IJISRT24OCT687

Ifft, J., Kuhns, R., & Patrick, K. (2018). Can machine learning improve prediction: An application with farm survey data. International Food and Agribusiness Management Review, 21(8), 1083–1098. https://doi.org/10.22434/IFAMR2017.0098

Key, N. (2022). The determinants of beginning farm success. Journal of Agricultural and Applied Economics, 54(2), 199–223. https://doi.org/10.1017/aae.2022.6

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Leanpub.

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63. https://doi.org/10.48550/arXiv.2010.16061

Reid, M. D., & Williamson, R. C. (2011). Information, divergence and risk for binary experiments. Journal of Machine Learning Research, 12, 731–817. https://doi.org/10.48550/arXiv.0901.0356

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sha D, Du P, Wu L. Classification and Prediction of Food Safety Policy Tools in China Based on Machine Learning. J Food Prot. 2024 Jun;87(6):100276. https://doi.org/10.1016/j.jfp.2024.100276

Shakoor, Md. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). Agricultural production output prediction using supervised machine learning techniques. 2017 1st International Conference on Next Generation Computing Applications (NextComp), 182-187. https://doi.org/10.1109/NEXTCOMP.2017.8016196

Szumigraj, A. (2022). The relationship between income and assets in farms and context of sustainable development. Problems of Agricultural Economics, 4(373), 5–24. https://doi.org/10.22004/ag.econ.329847

Zhang, Y., Zhao, Z., & Zheng, J. (2017). A comparison of machine learning algorithms for predicting student performance. Journal of Physics: Conference Series, 887, 012028. https://doi.org/10.1088/1742-6596/887/1/012028

Zhou, M., Liu, Q., Yang, T., Wu, Z., & Zhou, S. (2024). Impacts of innovation drivers in Chinese cities: A machine learning analysis using XGBoost. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5044198

# PACCHETTI UTILIZZATI

Bates, D., Maechler, M., & Dai, B. (2023). Matrix: Sparse and dense matrix classes and methods (R package version 1.6-0). https://CRAN.R-project.org/package=Matrix

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions (tree-shap package). Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77. https://doi.org/10.1186/1471-2105-12-77

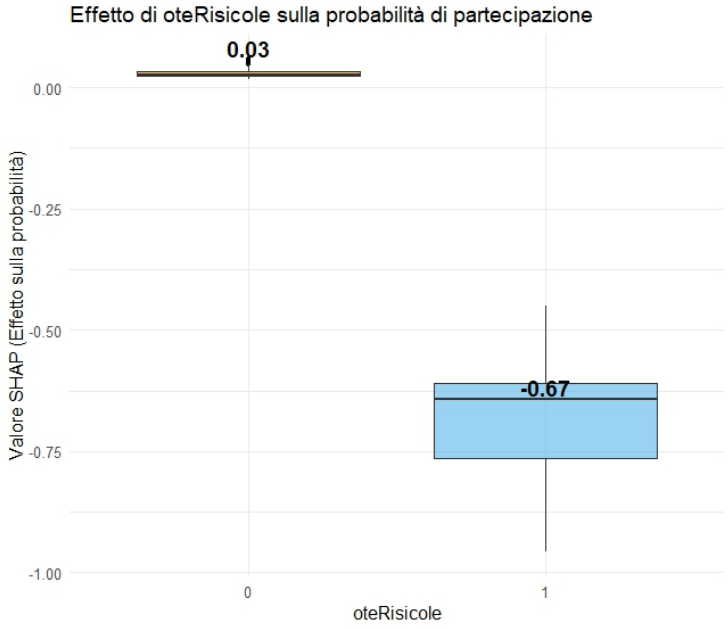
Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

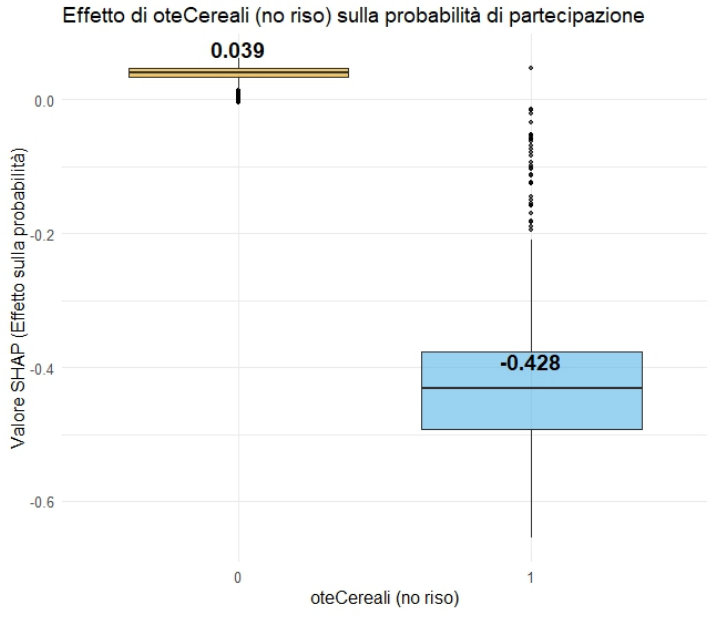
Wickham, H., & Henry, L. (2023). tidyr: Tidy messy data (R package version 1.3.0). https://CRAN.R-project.org/package=tidyr

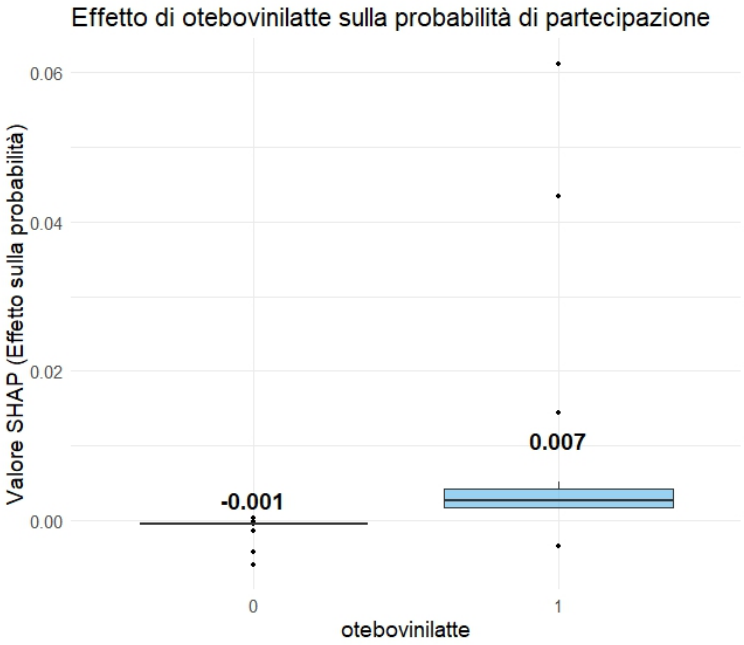
Wickham, H., François, R., Henry, L., & Müller, K. (2023). dplyr: A grammar of data manipulation (R package version 1.1.3). https://CRAN.R-project.org/package=dplyr

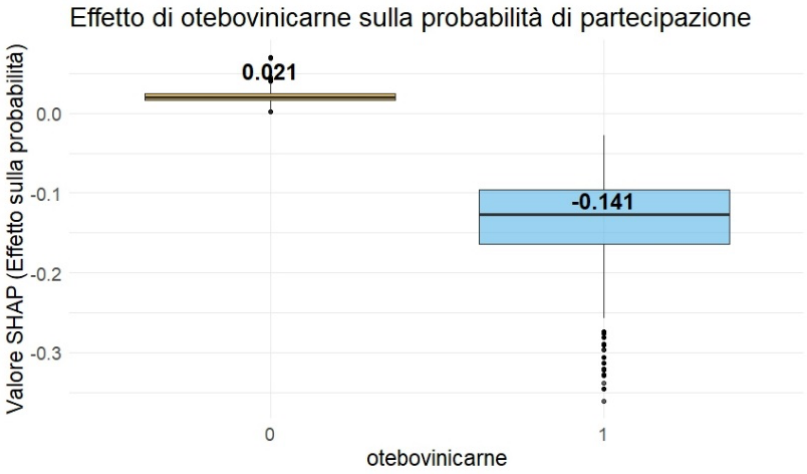
# ALLEGATI

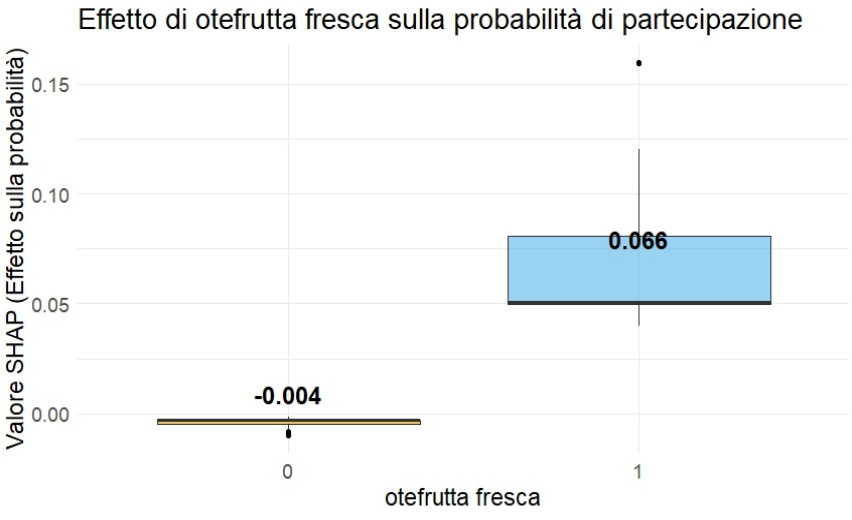
## ALLEGATO 1– DEPENDENCE PLOT DEGLI ORIENTAMENTI TECNICO ECONOMICI DESCRITTI AL PARAGRAFO 5.3.5

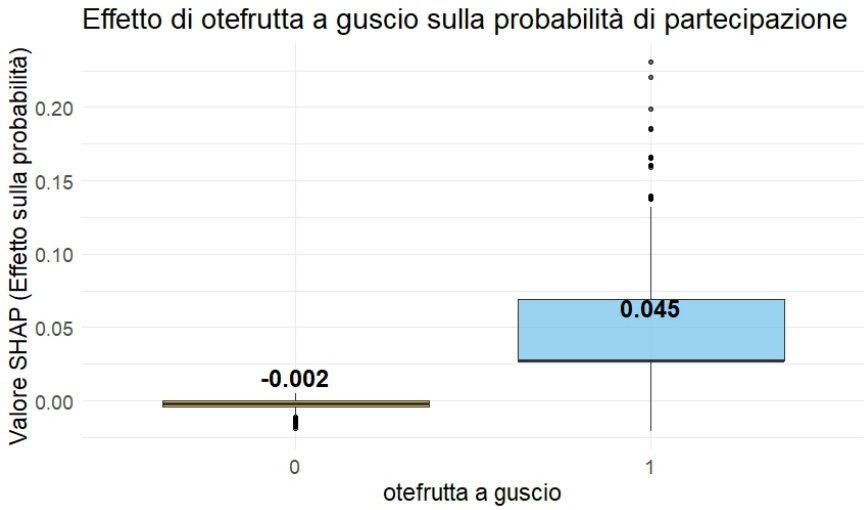


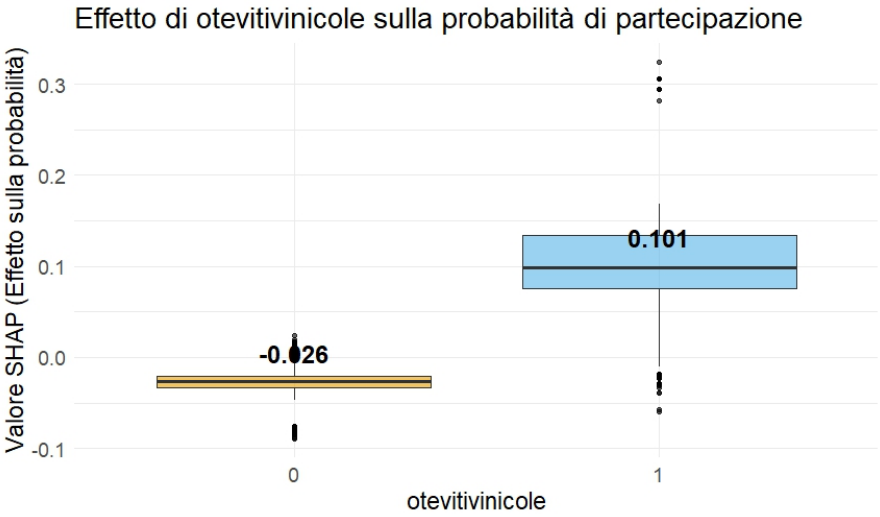




Effetto di otebovinilatte sulla probabilità di partecipazione
Valore SHAP (Effetto sulla probabilità)
0.06
0.04
0.02
0.00
-0.001
0.007
0
1
otebovinilatte


Effetto di otebovinicarne sulla probabilità di partecipazione
Valore SHAP (Effetto sulla probabilità)
0.0
-0.1
-0.2
-0.3
0.021
-0.141
0
1
otebovinicarne


Effetto di otefrutta fresca sulla probabilità di partecipazione
Valore SHAP (Effetto sulla probabilità)
0.15
0.10
0.05
0.00
0.066
-0.004
0
1
otefrutta fresca


Effetto di otefrutta a guscio sulla probabilità di partecipazione
Valore SHAP (Effetto sulla probabilità)
0.20
0.15
0.10
0.05
0.00
-0.002
0.045
0
1
otefrutta a guscio

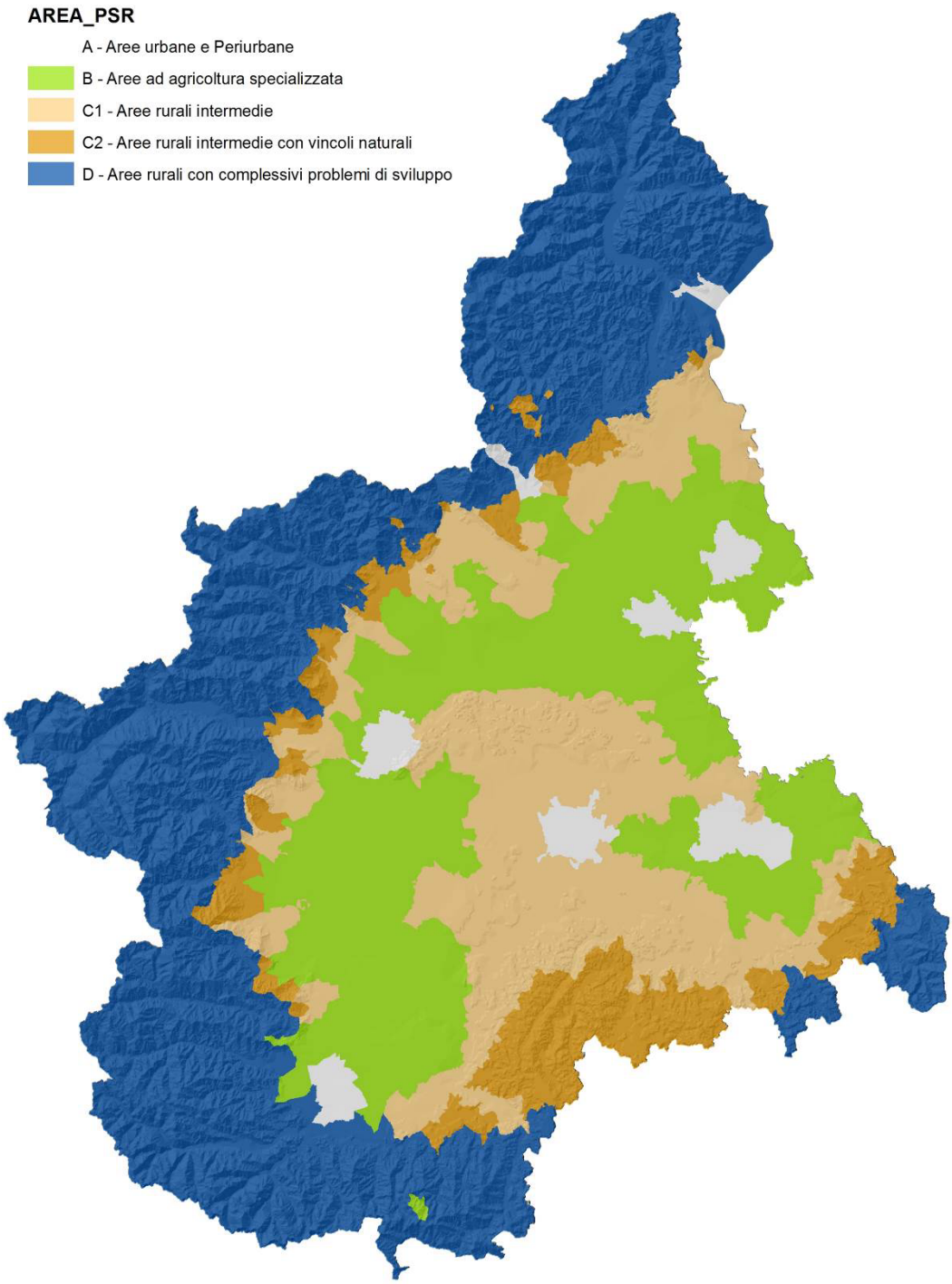

Effetto di otevitivinicole sulla probabilità di partecipazione
Valore SHAP (Effetto sulla probabilità)
0.3
0.2
0.1
0.0
-0.1
-0.026
0.101
0
1
otevitivinicole

# ALLEGATO 2 – DEPENDENCE PLOT PER LA VARIABILE TERRITORIALE DESCRITTA AL PARAGRAFO 5.3.6

La variabile sulla localizzazione fa riferimento alla classificazione adottata dal PSR che individua 5 diverse tipologie territoriali:

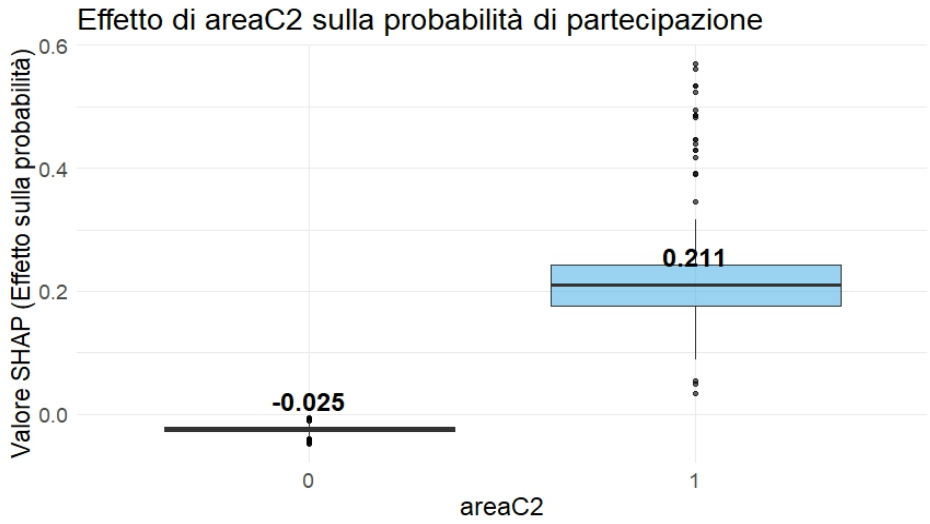
- Area A – Poli Urbani: in questa tipologia sono inseriti solo gli otto capoluoghi delle Province Piemontesi.
- Area B – Area ad agricoltura specializzata. sono quelle di pianura. In queste aree ricadono non solo i comuni prettamente agricoli, ma anche tutte le aree periurbane con agricoltura interstiziale ad alta densità di popolazione.
- Area C1 – Aree Intermedie. Si estendono sulle zone collinari del Piemonte. Queste aree comprendono, quindi tutte le aree più vocate alla viticoltura, tra cui spicca la zona delle langhe nel quale si è innescato in tempi relativamente recenti un processo di sviluppo turistico decisamente interessante.
- Aree C2 – Aree intermedie con vincoli naturali. Sono rappresentate dalla fascia di comuni che si trovano ai piedi delle Alpi, sull'Appennino o nell'Alta Langa. Queste aree sono state separate dalle aree C1 in quanto le pendenze medie ed i principali indicatori socio economici dimostrano che queste sono aree problematiche.
- Aree D – Aree rurali con complessivi problemi di sviluppo. Si estendono sulle aree montane. Al loro interno presentano livelli di sviluppo più disomogenei rispetto alle Aree C2, in quanto diversi comuni classificati come D si trovano o in località sciistiche rinomate (alta Val Susa, Val Chisone, Area del Rosa, Limone Piemonte) oppure sono limitrofi alla zona del Lago Maggiore, che è una delle località regionali più vocate al turismo. I comuni estranei a queste aree più sviluppate, in particolare quelli localizzati tra i 650-700 ed i 1000 metri, invece hanno difficoltà sociali ed economiche più marcate.

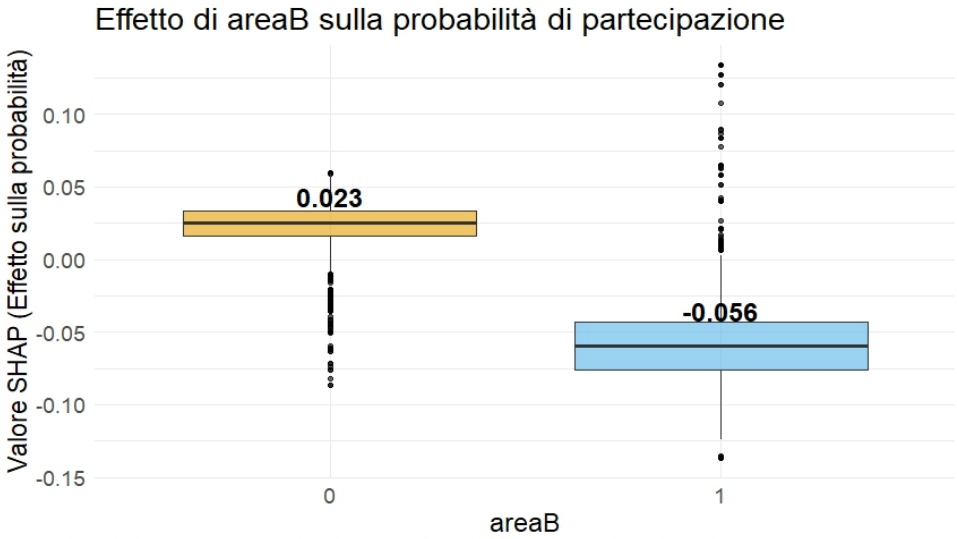
**Figura 7 - Le tipologie territoriali del PSR 2014 – 2020 del Piemonte**

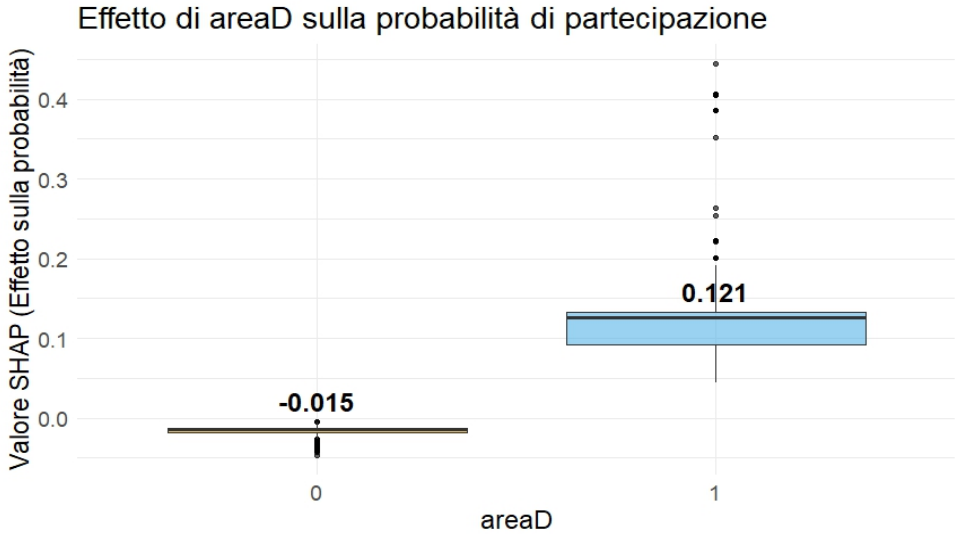



Fonte: Elaborazione Ires Piemonte su dati Regione Piemonte

I plot sottostanti si riferiscono solo ai territori per i quali era maggiore il valore SHAP medio assoluto.



# ALLEGATO 3 – SCRIPT UTILIZZATO PER L'ANALISI

Di seguito si riporta lo script utilizzato per l'analisi oggetto di questo rapporto. È stato utilizzato l'algoritmo XGBoost per costruire un modello predittivo e SHAP (SHapley Additive ExPlanations) per interpretare i risultati, identificando le variabili più influenti e analizzandone gli effetti sulla probabilità di partecipazione. Lo script è suddiviso in diverse sezioni: (1) caricamento delle librerie e dei dati, (2) pre-elaborazione dei dati, (3) ottimizzazione degli iperparametri e addestramento del modello XGBoost, (4) valutazione del modello, (5) interpretazione tramite i valori SHAP, e (6) creazione di grafici per visualizzare l'effetto delle variabili e delle loro interazioni. Il tutto è stato eseguito in R Studio (versione 2024.12.1 build 563)."

```
# Caricamento librerie

library(xgboost)
library(treeshap)
library(dplyr)
library(ggplot2)
library(caret)
library(tidyr)
library(pROC)
library(Matrix)

# Caricamento dati
df <- DB_xgboost_tutti_i_bandi
# Conversione variabile target in binaria (0 = NO, 1 = SI)
df$partecipante <- ifelse(df$partecipante == "SI", 1, 0)
# Gestione variabili categoriche
df$ote <- as.factor(df$ote)
df$area <- as.factor(df$area)
df$sesso <- as.factor (df$sesso)
# Separazione train/test set
set.seed(123)
trainIndex <- createDataPartition(df$partecipante, p = 0.8, list = FALSE)
train <- df[trainIndex, ]
test <- df[-trainIndex, ]
# Conversione in matrice numerica per XGBoost
train_matrix <- model.matrix(partecipante ~ . -1, data = train)
test_matrix  <- model.matrix(partecipante ~ . -1, data = test)
# Creazione dei DMatrix per XGBoost
dtrain <- xgb.DMatrix(data = train_matrix, label = train$partecipante)
dtest  <- xgb.DMatrix(data = test_matrix, label = test$partecipante)
#XGBOOST----
#-----MIGLIORI IPER PARAMETRI-------
dfpar <- DB_xgboost_tutti_i_bandi
dfpar$partecipante <- factor(dfpar$partecipante, levels = c("NO", "SI"))
# Definizione della griglia di iperparametri
xgb_grid <- expand.grid(
  nrounds = c(50, 100, 200),          # Numero di alberi
  max_depth = c(3, 5, 7),             # Profondita massima degli alberi
  eta = c(0.01, 0.1, 0.3),            # Tasso di apprendimento
  gamma = c(0, 1, 5),                 # Penalizzazione per foglie inutili
  colsample_bytree = c(0.6, 0.8, 1),  # Frazione di colonne usate per albero
  min_child_weight = c(1, 3, 5),      # Peso minimo richiesto in un nodo
  subsample = c(0.6, 0.8, 1)          # Percentuale di dati usati per ogni albero
```

```
)
# Controllo della validazione incrociata
ctrl <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)

# Addestramento con ricerca dei migliori iperparametri
xgb_model <- train(
  partecipante ~ .,
  data = dfpar,
  method = "xgbTree",
  tuneGrid = xgb_grid,
  trControl = ctrl,
  metric = "ROC"
)

# Stampa i migliori parametri trovati
best_params <- xgb_model$bestTune
print(best_params)

#---MODELLO XGBOOST------

#Dati sbilanciati: calcolo la class_ratio
class_ratio <- sum(train$partecipante == 0) / sum(train$partecipante == 1)
print(class_ratio)
params <- list(
  objective = "binary:logistic",
  eval_metric = "logloss",
  max_depth = 3,
  eta = 0.1,
  subsample = 0.8,
  colsample_bytree = 1,
  scale_pos_weight = class_ratio,
  min_child_weight = 3)
# Training del modello
set.seed(123)
xgb_model <- xgb.train(
  params = params,
  data = dtrain,
  nrounds = 200,
  watchlist = list(train = dtrain, test = dtest),
  early_stopping_rounds = 10)

# Predizioni e valutazione
pred_probs <- predict(xgb_model, dtest)
pred_labels <- ifelse(pred_probs > 0.5, 1, 0)
auc_value <- roc(test$partecipante, pred_probs)$auc
conf_matrix <- confusionMatrix(factor(pred_labels), factor(test$partecipante))
f1_score <- conf_matrix$byClass["F1"]
# Stampa metriche
cat("AUC-ROC:", auc_value, "\n")
cat("F1-score:", f1_score, "\n")
# Importanza delle variabili
xgb_importance <- xgb.importance(feature_names = colnames(train_matrix), model = xgb_model)
# Plot importanza XGBoost
xgb.ggplot.importance(xgb_importance)
```

```
#--------INTERPRETAZIONE CON SHAP (TREESHAP) -------------------------------

# Estraggo i dati originali usati per creare dtrain
train_matrix <- as.data.frame(train_matrix)
# Aggiungo i nomi delle colonne per garantire compatibilità
colnames(train_matrix) <- colnames(model.matrix(partecipante ~ . -1, data = train))
# Unificazione modello con treeshap
unified_model <- xgboost.unify(xgb_model, train_matrix)
# Calcolo SHAP
test_matrix <- as.data.frame(test_matrix)
colnames(test_matrix) <- colnames(model.matrix(partecipante ~ . -1, data = test))
#tiro fuori le SHAP VALUES
shap_values <- treeshap(unified_model, test_matrix)


#----CALCOLO LA BASELINE LOGIT e tabella con valori shap e probabilita---------
pred_logit <- predict(xgb_model, dtrain, outputmargin = TRUE)
baseline_logit <- mean(pred_logit)
cat("Baseline logit:", baseline_logit, "\n")
logit_to_prob <- function(logit) {
  return(1 / (1 + exp(-logit)))
  }
shap_to_prob <- function(shap_value, baseline_logit) {
  logit <- baseline_logit + shap_value
  # Calcolare la probabilità usando la funzione sigmoide
  prob <- logit_to_prob(logit)
  return(prob)
}
shap_values_seq <- seq(-2, 2, by = 0.2)
prob_table <- data.frame(
  SHAP_Value = shap_values_seq,
  Probability = sapply(shap_values_seq, shap_to_prob, baseline_logit = baseline_logit)
)
print(prob_table)
# Estraggo i valori SHAP come dataframe
shap_values_df <- as.data.frame(shap_values$shaps)
# Aggiungo i nomi delle feature per chiarezza
colnames(shap_values_df) <- colnames(test_matrix)
# aggiungo le osservazioni originali per un confronto
shap_values_df$Observation <- rownames(test_matrix)
# Calcolo l'importanza media assoluta delle feature
shap_importance <- data.frame(
  Feature = colnames(shap_values_df)[-ncol(shap_values_df)],  # Escludi la colonna "Observation"
  Mean_Abs_SHAP = colMeans(abs(shap_values_df[, -ncol(shap_values_df)]))  # Media assoluta dei valori
SHAP
)
# Ordino per importanza e plotto
ggplot(shap_importance, aes(x = reorder(Feature, Mean_Abs_SHAP), y = Mean_Abs_SHAP)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Importanza delle Variabili (SHAP)", x = "Variabile", y = "SHAP Value Medio Assoluto")

#-------DEPENDENCE PLOT ---------------------------------------------------


#Plot con VARIABILI CATEGORIALI ------------
# Seleziono la variabile categoriale
feature <- "oteCereali (no riso)"
# Creo il dataframe con valori SHAP
shap_df <- data.frame(
  Feature_Value = as.factor(test_matrix[, feature]),  # Converti in fattore
```

```
  SHAP_Value = shap_values$shaps[, feature]  # Valori SHAP
)

# Calcolo la media SHAP per ciascun gruppo
shap_means <- shap_df %>%
  group_by(Feature_Value) %>%
  summarise(mean_shap = mean(SHAP_Value), .groups = "drop")
# Plot con media SHAP e colori distintivi
ggplot(shap_df, aes(x = Feature_Value, y = SHAP_Value, fill = Feature_Value)) +
  geom_boxplot(alpha = 0.6, outlier.color = "black", outlier.size = 1) +
  geom_text(data = shap_means, aes(x = Feature_Value, y = mean_shap,
                                   label = round(mean_shap, 3)),
            vjust = -1, size = 5, fontface = "bold") +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +  # Colori distintivi
  labs(title = paste("Effetto di", feature, "sulla probabilità di partecipazione"),
       x = feature,
       y = "Valore SHAP (Effetto sulla probabilità)") +
  theme_minimal() +
  theme(legend.position = "none",
        text = element_text(size = 14))

#----Plot per VARIABILE CONTINUA---------
# scelgo la feature

feature <- "ps"

#df con valori SHAP per feature

shap_df <- data.frame(
  Feature_Value = test_matrix[, feature],
  SHAP_Value = shap_values$shaps[, feature],
  Interaction = test_matrix[, "ps"]
)
#PLOT per variabile continuna
ggplot(shap_df, aes(x = Feature_Value, y = SHAP_Value, color = as.factor(Interaction))) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", color = "black") +
  labs(title = paste("Effetto della", feature, "sulla probabilità di partecipazione"),
       x = feature,
       y = "Valore SHAP") +
  theme_minimal() +
  theme(legend.position = "none")+
  xlim(0,2000) # Aggiunta di xlim per limitare l'asse x

#----CREAZIONE DEPENDENCE PLOT INFLUENZA VARIABILE SU SHAP DI ALTRA VARIABILE—


# Estraggo i nomi delle colonne relative a 'ote' dai valori SHAP
ote_cols <- grep("^ote", colnames(shap_values$shaps), value = TRUE)
# Converto i valori SHAP in formato long per 'ote'
shap_ote_long <- shap_values$shaps %>%
  as.data.frame() %>%
  select(all_of(ote_cols)) %>%
  mutate(Observation = 1:nrow(.)) %>%
  tidyr::pivot_longer(cols = all_of(ote_cols), names_to = "OTE_Mode", values_to = "SHAP_Value") %>%
  mutate(OTE = sub("^ote", "", OTE_Mode))  # Rimuovo il prefisso "ote" per allinearlo a 'test$ote'
# Unisco con il dataset originale 'test' per avere 'ps' e 'ote'
df_plot <- test %>%
  select(ps, ote) %>%
  mutate(Observation = 1:nrow(.)) %>%  # Aggiungo un ID per il join
  left_join(shap_ote_long, by = "Observation") %>%
  filter(OTE == ote) %>%  # Filtro solo le righe dove la categoria SHAP corrisponde a 'ote'
```

```
  select(ps, ote, SHAP_Value)  # Mantengo solo le colonne rilevanti
# Filtro categorie di interesse
df_plot_filtered <- df_plot %>%
  filter(ote %in% c("Risicole"))

# Verifico quante osservazioni ci sono per categoria
print(table(df_plot_filtered$ote))

# Grafico separato per categoria------ usare “facet wrap”
ggplot(df_plot_filtered, aes(x = ps, y = SHAP_Value, color = ote)) +
  geom_point(alpha = 0.3, size = 1.5) +
  geom_smooth(method = "loess", se = FALSE, linewidth = 1.5) +
  scale_x_continuous(limits = c(0, 500), breaks = seq(0, 500, by = 50)) +
  scale_color_manual(values = c("frutta fresca" = "#E69F00", "frutta a guscio" = "#56B4E9")) +
  labs(title = "Effetto di PS sui valori SHAP di OTE",
       x = "Produzione Standard (in .000 di euro)",
       y = "Valore SHAP di OTE") +
  facet_wrap(~ ote, scales = "free_y") +
  theme_minimal() +
  theme(legend.position = "none",
        text = element_text(size = 12),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

NOTE EDITORIALI

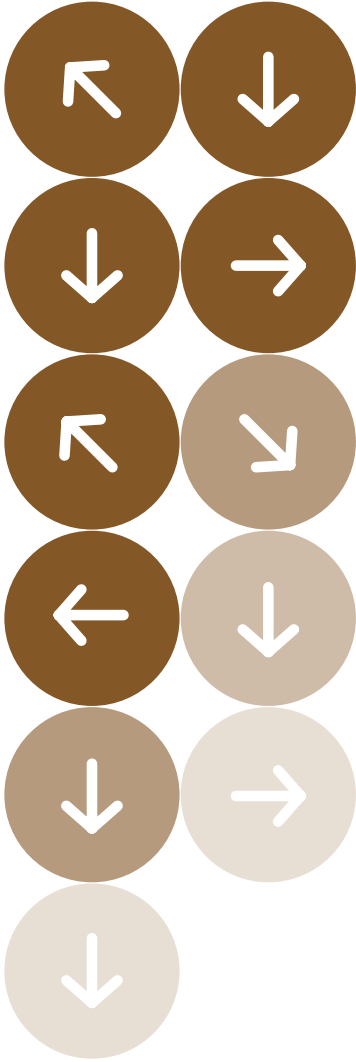
Editing
IRES Piemonte

Ufficio Comunicazione
Maria Teresa Avato


Luglio 2025
Istituto di Ricerche Economico Sociali del Piemonte
Via Nizza 18 -10125 Torino

www.ires.piemonte.it

*Ambiente e Territorio*
*Cultura*
*Finanza locale*
*Immigrazione*
*Industria e Servizi*
*Istruzione e Lavoro*
*Popolazione*
*Salute*
*Sviluppo rurale*
*Trasporti*



**IRES Piemonte**
Via Nizza, 18
10125 TORINO
+39 0116666-461
www.ires.piemonte.it